Contents

1 Getting Data

1.1 Exploratory Data Analysis

Definition 1.1.1 ▸ Population

A **population** is the entire group (of individuals or objects) that we wish to know something about.

Definition 1.1.2 ▸ Research Question

A **research question** is usually one that seeks to investigate some characteristic of a population.

Briefly speaking, we can classify **research questions** into the following categories

1. To make an estimate about the population
2. To test a claim about the population
3. To compare two sub-populations / to investigate a relationship between two variables in the population

Definition 1.1.3 ▸ Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a systematic process where we explore a data set and its variables and come up with **summary statistics** as well as **plots**. EDA is usually done **iteratively** until we find useful information that helps us answer the questions we have about the data set.

In general, the steps involved in EDA are:

1. Generate **research questions** about the data.
2. Search for answers to the research questions using data visualization tools. In the process of exploration, we could also perform data modeling. (e.g. regression analysis)
3. We ask ourselves the following question: "To what extent does the data we have, answer the questions we are interested in?"

4. We refine our existing questions or generate new questions about the data before going back to data for further exploration.

1.2 Sampling

The overall workflow for sampling is as follows (**This is very important**!)

1. **Define Population of Interest**: Clearly identify the group you want to study.
2. **Establish Sampling Frame**: Create or obtain a comprehensive list from which to select your sample.
3. **Evaluate Sampling Frame Quality**: Check for issues such as:
    - Coverage (does it include all population elements?)
    - Relevance (does it contain extraneous units?)
    - Duplication (are some units listed multiple times?)
    - Clustering (are units grouped in a way that requires special sampling approaches?)
4. **Choose a Sampling Method and select sampling units**: Choose an appropriate technique (e.g., simple random, stratified, cluster sampling) based on your sampling frame characteristics.
5. **Collect Data**: Gather the required **information** from your **selected sample**.
6. **Calculate Estimates**: Use sample data to make inferences about the population parameters.
7. **Assess Precision and Bias**: Evaluate the reliability of your estimates.
8. **Make Inferences**: Draw conclusions about the population based on your sample data.

1.2.1 Basic Concepts and Definitions

Definition 1.2.1 ▶ Population of Interest

A **population of interest** refers to a group in which we have interest in drawing conclusions in a study.

Definition 1.2.2 ▶ Population Parameter

A **population parameter** is a **numerical** fact about a population.

Example 1.2.3 ▶ Example of Population Parameters and Population

The following are some examples of a population and an associated population parameter.

1. The average height (population parameter) of all primary six students in a particular primary school (population).
2. The median number of modules taken (population parameter) by all first-year undergraduates in a University (population).
3. The standard deviation of the number of hours spent on mobile games (population parameter) by pre-schoolers aged 4 to 6 in Singapore (population).

Definition 1.2.4 ▶ Census

A **census** is a **method of data collection** from everyone. It attempts to reach out to the entire population of interest.[a]

[a] Note that since a census does not involve sampling, the notion of sampling frame is not applicable in that context.

Some disadvantages of conducting a census are as follows:

1. Conducting a census is often prohibitively expensive.
2. Some studies require timely data, but censuses typically take a long time to complete.
3. Even when attempted, a census may not achieve a 100% response rate.

Definition 1.2.5 ▶ Sampling

As contrast to **census**, if we don't reach out to everyone from the entire population of interest, we only select part of them using the sampling methods we will introduce later, this **method of data collection** is called **sampling**.

In **Sampling**, we have the following important three concepts:

1. **Sampling Frame**[1]: A **sampling frame** is the list from which the **sample** was obtained.
2. **Sample**[2]: A **sample** is a **proportion of the population** selected in the study. It's fundamental element is called **sampling unit** and it is selected from the **sampling frame**.

[1] The **sampling frame** provides **a list of samples** we can select our sample from.
[2] Doing so is because it is usually not feasible to gather information from every member of the population

3. **Estimate**: An **Estimate** is an inference about the **population parameter** based on the information obtained from a **sample**.

In EDA, the most important question is whether the sample obtained from such a sampling frame is still **able to tell us something about the population parameter**. Regarding this, the following are some important characteristics of **sampling frame**:

1. Does the sampling frame include **all available sampling units**[3] from the population?
2. Does the sampling frame contain **irrelevant** or **extraneous** sampling units from another population?
3. Does the sampling frame contain **duplicated** sampling units?
4. Does the sampling frame contain sampling units in **clusters**?

In our sampling frame, we are more interested in the **generalisability**, which is the ability to generalise the findings from a sample to the population. And the requirement is that the **sampling frame must be equal to or greater than the population of interest**.[4]

Population of Interest vs. Sampling Frame vs. Sample

1. **Population of Interest**
    - The complete set of individuals or units you want to study and make conclusions about
    - Example: All people who drink coffee in Singapore
2. **Sampling Frame**
    - The operational list or source from which you actually select your sample
    - It's your practical way to access the population
    - Example: A customer database from coffee shops, a residential directory, or registered participants in a consumer survey who indicated they drink coffee
3. **Sample**
    - The subset of units actually selected from the sampling frame for data collection
    - These are the specific individuals or units you'll gather information from
    - Example: The 500 coffee drinkers you ultimately survey or interview

The relationships between them:

[3] Sampling units are the individual elements or entities that are selected from a population when conducting a survey or research study.

[4] This means that the sampling frame should include **every sampling unit** from the population of interest.

- Ideally, your sampling frame would perfectly match your population of interest, but this rarely happens in practice
- Your **sample is drawn directly from your sampling frame**, not from the theoretical population
- Any limitations in your sampling frame (like missing certain types of coffee drinkers) will affect how well your sample represents your population of interest

1.2.2 Bias

Bias refers to systematic errors that skew sample data, preventing it from accurately representing the population and leading to unreliable conclusions. One consequence is that, even if our **sampling frame** covers the entire population of interest, our findings from the sample **may not always be generalisable** to the population. There are **two** major kinds of biases.

1. **Selection Bias**: Occurs when the researcher's **biased selection of units** – due to an imperfect sampling frame or non-probability sampling – excludes certain units, preventing the sample from representing the population accurately.
2. **Non-response Bias**: Arises from **participants' non-participation or non-disclosure**, excluding their information due to reasons like inconvenience or sensitivity, regardless of the sampling method used.

> *Remark.* The **bias** will affect the construction of our **sampling frame**, and since our **sample units** are chosen from the sampling frame to form the **sample we want to gather information from**, so it will affect our **generalisability** from sample to population.

1.2.3 Probability Sampling

Definition 1.2.6 ▶ dfn-probability-sampling

Probability sampling is a sampling scheme such that the selection process is done via a known randomised[a] mechanism.[b]

[a] The randomized mechanism is important as it introduces an element of chance in the selection process so as to **eliminate biases**.

[b] It is important that every unit in the sampling frame has a **known non-zero probability** of being selected but the probability of being selected does not have to be same for all the units.

We will introduce four main types of probability sampling methods:

Simple Random Sampling

Simple **R**andom **S**ampling, or SRS, is a method where every possible group of n units from the population has an equal chance of being selected. For example, to choose 5 students from a class of 30, assign each student a number from 1 to 30 and use a **random number generator** to pick 5 distinct numbers.

1. SRS is usually achieved by using a **random number generator**. However, one **shortcoming** for SRS is that it can possibly subjected to **non-response** from the units that are sampled.
2. SRS needs to know the number **whole sampling unit in the population**! Otherwise, you can't ensure that you can't ensure that every group of n has an equal probability of selection. Alternatively speaking, when you use your calculator, like CASIO 991 to implement a random number generator, you need to specify the range of your numbers (you total sampling unit).

Systematic Sampling

Systematic sampling is a method of selecting units from a **list**[5] by applying a selection interval k and a **random starting point** from the **first interval**. To carry out systematic sampling:

1. Suppose we know how many sampling units there are in the population (denoted by p)
2. We decide how big we want our sample to be (denoted by n). This means that we will select one unit from every $k = \frac{p}{n}$ units;
3. From 1 to $k = \frac{p}{n}$, select a number **at random**, say r
4. With this, the sample will consist of the following units from the list:

$$r, r+k, r+2k, \cdots, r+(n-1)$$

However, it is **often** that we do not know the number of sampling units p in the population. In such a situation, systematic sampling can still be done by **deciding on the selection interval** k and randomly selecting a unit from the first k units and then subsequently every kth unit will be sampled. For example, if $k = 10$, we can sample the $5^{\text{th}}, 15^{\text{th}}, 25^{\text{th}}$ units and so on.

[5] the "list" refers to the sampling frame, which is a complete list of all units in the population that you can sample from.

1. Compared to SRS, systematic sampling doesn't need to know the whole sampling units in the sampling frame.
2. The shortcome is that, if the **list**, or **sampling frame**, is **not random**, but instead contains some inherent grouping or ordering of the units, then the sample generated by **systematic sampling** may not be **representative** of the population.

Stratified Sampling

Stratified Sampling is a sampling method that combines the idea of "classification" and "randomization" and it is achieved as follows:

1. The sampling frame is divided into groups called **strata**. Each **stratum** is **similar in that they share similar characteristics** but the size of each stratum does not necessarily have to be the same.
2. SRS is then applied to each stratum to generate the whole sample.

1. Because of the property of **stratum**, which is within each stratum, the sample units must share similar characteristic, it is sometimes hard to form such stratum.

Cluster Sampling

It is similar to Stratified Sampling, but the difference lies in **how the whole sampling frame is classified**

1. Divide the sampling frame into **clusters**, where **clusters** doesn't have any requirements for its inner sampling units, thus ensuring the **inner diversity**
2. Use SRS to select **a fixed number of clusters**[6].
3. All the sampling units from the selected clusters are then included in the overall sample.

Table 1.1 summarizes the advantages and disadvantages for the four probability sampling methods we have learned

[6]Now sampling units within each cluster here

Sampling Plan	Advantages	Disadvantages
Simple Random Sampling	Good representation of the population	Time-consuming; accessibility of information and sampling frame
Systematic Sampling	Simple selection process as opposed to simple random sampling	Potentially under-representing the population
Stratified Sampling	Good representation of the sample by stratum	Require sampling frame and criteria for classification of the population into stratum
Cluster Sampling	Less time-consuming and less costly	Require clusters to be reasonably heterogeneous[7] and not have cluster-specific characteristics

Table 1.1: Comparison of Sampling Plans

1.2.4 Non-probability Sampling

Definition 1.2.7 ▶ dfn-non-probability-samping

non-probabilty sampling method is when the selection of units is **not done by randomisation**.

Convenience Sampling

Convenience sampling is when a researcher chooses the sampling units to form a sample among those that are most easily available to participate in the study.

Volunteer Sampling

Volunteer sampling is when the sampling units **volunteer** themselves into a sample.

1.3 Variables and Summary Statistics

1.3.1 Variables

Let's kick start this part by introducing some concepts

1. A **variable** is an attribute that can be measured or labelled.
2. A **data set** is a collection of **individuals** and **variables pertaining to the individuals**. Individuals can refer to either objects or people.

In a research question where we are examining relationships between variables, there is usually a distinction between which are **independent variables** and which are **dependent**

variables.

> *Remark.* It is important to note that the dependent variable is **hypothesised to change** when the independent variable is manipulated. It does not mean that the dependent variable **must** change.

Types of Variables

1. **Categorical Variable**: Variables that take on categories or label values. These categories or labels are **mutually exclusive**.
    - **Ordinal Variable**: A categorical variable where there is some natural ordering and numbers can be used to represent the ordering.
    - **Nominal Variable**: A categorical variable where there is no intrinsic ordering
2. **Numerical Variable**: Variables that take on numerical values and we are able to **meaningfully perform arithmetic operations** like adding and taking average.
    - **Discrete Variable**: A numerical variable where there are gaps in the set of possible numbers taken on by the variable. For example, the number of students can only be **integer**, cannot have fraction.
    - **Continuous Variable**: A numerical variable that can take on all possible numerical values in a given range or interval. For example, the height of a student can be any number, not restricted to integer only.

1.3.2 Summary Statistics

Summary Statistics can be divided into two parts

1. Those that measure **central tendencies** of the data, like **mean, median** and **mode**
2. Those that measure the **level of dispersion (or spread)** of the data, like **standard deviation** and **interquartile range**.

Measure Central Tendencies

Mean

The mean is simply the average value of a numerical variable x. We denote the mean of x by $\bar{x}$ and the formula to compute $\bar{x}$ is

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n} = \frac{\Sigma_{i=1}^{n} x_i}{n}$$

There are some properties of the **mean** of a variable

1. Even if we don't know each of the individual value, we can still calculate the sum, given the value of $\bar{x}$ and the number of data points n, by using sum $= \bar{x} \times n$
2. Adding a constant value c to all the data points changes the mean by that constant value. For example, if we add a constant c to each of the data point, the new mean will be $\bar{x} + c$
3. Multiplying a constant value of c to all the data points will result in the new mean being changed by the same factor of c. For example, if we multiply a constant c to each of the data point, the new mean will be $c \times \bar{x}$

Median

The median is simply the middle value of the variable after arranging the values of the data set in ascending or descending order. The meaning of **median** is that there are 50% data points which are below the median, and correspondingly, there are 50% data points which are above the median.

> *Remark.* If there are two middle values (when there are an even number of data points), we will take the **average** of the two middle values as the median.

There are some properties of **median** also

1. Adding a constant value c to all data points will result in the new median being **added** by the constant c.
2. Multiplying a constant value c to all data points will result in the new median being **multiplied** by the constant c.

Mode

The mode is a variable that appears **most often** in data.

> *Remark.* Mode is applicable to both **numerical** and **categorical** variables.

Measure Spread

Standard Deviation

Standard deviation is one of the ways to measure the **spread** of the data about the **mean** (denoted by $\bar{x}$). It can be computed as follows

$$\text{Sample Variance, Var} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

$$\text{Standard Deviation, } s_x = \sqrt{\text{Var}}$$

> *Remark.* The reason for the denominator to be $n-1$ not n is beyond the scope of this course.

There are some properties of the **standard deviation** of the data

1. The standard deviation s_x is always **non-negative**. Only when all the data points are of the same value then $s_x = 0$
2. The standard deviation shares **the same unit** as the numerical variable x.
3. Adding a constant c to all data points **doesn't change** the standard deviation.
4. Multiplying a constant c to all data points results in the standard deviation being multiplied by $|c|$, which is the absolute value of c.

Interquartile Range (IQR)

As we have seen above about the meaning of median, median is also called the 50^{th} percentile of the data values. Similarly, we have

1. **The first quartile**, denoted by Q_1, is the 25^{th} percentile of the data values.
2. **The third quartile**, denoted by Q_3, is the 75^{th} percentile of the data values.
3. **The Interquartile Range (IQR)**, which is $Q_3 - Q_1$

Similarly, there are some properties about the IQR

1. IQR is **non-negative**
2. Adding a constant c to all data points **doesn't change** the IQR.
3. Multiplying a constant c to all data points results in the IQR being multiplied by c

To find/calculate the Q_1 and Q_3 of a variable, given that the number of data points is n, we can use

1. **Find the median** of the total n data points.
2. **Divide the data into upper half and lower half** according to the median
   - If n is **even**, just divide normally
   - If n is **odd**, **exclude** the median from the upper half
3. **Find** Q_1 and Q_3
   - Q_1 is the median of the lower half.
   - Q_3 is the median of the upper half.

2

Categorical Data Analysis

2.1 Rates

Definition 2.1.1 ▶ Rates

In categorical data analysis, a **rate** (also called a proportion or relative frequency) quantifies how frequently a specific category occurs within a dataset. It is calculated as the ratio of the number of observations in a **given category** (a.k.a **target**) to the **total** number of observations (a.k.a **population**). **Rates** are typically expressed as a fraction (between 0 and 1) or a percentage (between 0% and 100%), providing a standardized measure of prevalence or likelihood for that category.

Remark. Based on the fact the **rate** are always between 0 and 1 (or 0% and 100%), we can intuitively see that **rate** can be used as a fair comparison when group sizes are unequal. This idea is known as "changing from absolute numbers to percentage".

Example 2.1.2 ▶ One Variable Rate Example

A survey asks 500 people, "Do you prefer tea or coffee?" Of the respondents, 200 select "Tea" and 300 select "Coffee".

$$\text{rate(tea)} = \frac{200}{500} = 0.4(\text{or } 40\%)$$
$$\text{rate(coffee)} = \frac{300}{500} = 0.6(\text{or } 60\%)$$

Remark. For every **rate** problem, we need to find out what is our **population**, this is become the denominator. And what is our **target**, this will become our nominator.

Before we go deeper into the three different kinds of rates, let's give out the following table,

Remark. This table is known as a 2×2 **contingency table**. And the dependent variable *Outcome* is placed on the columns on the table while the independent variable *Treatment* is placed on the rows.

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

Table 2.1: Kidney Problem

2.1.1 Marginal Rate

Definition 2.1.3 ▸ Marginal Rate

Marginal rate, as its name suggests, are calculated by two **marginal** numbers in the table.

Example 2.1.4 ▸ Marginal rate Example

In the table 2.1 given above, suppose we want to know what proportion of total number of patients were given treatment Y and the proportion of total number of patients were given **success** treatment.

Solution: This problem is actually a **marginal rate** problem. And the two marginal data we need are labeled as red as follows:

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	**350**
Column Total	**831**	219	**1050**

Figure 2.1: Kidney Problem (Marginal rate)

So, we can calculate our rate(Y)[a] as follows:

$$\text{rate(Y)} = \frac{350}{1050} = 33\frac{1}{3}\%$$

Similarly, rate(X)[b] is also a marginal rate and will be

$$\text{rate(X)} = \frac{831}{1050} = 66\frac{2}{3}\%$$

[a] "Y" represents the patients who are given treatment Y
[b] "X" represents the patients who are given successful treatments

2.1.2 Conditional Rate

Definition 2.1.5 ▸ Conditional rate

Conditional rate, as its name suggests, must set a **condition** for the calculation of the rate.

Example 2.1.6 ▸ Conditional rate example

For example, with Table 2.1, if we want to know *Among those patients given treatment X, what proportion were successful?*.
Solution: This is a **conditional rate** problem. The condition specifies our **population** (X) to be 700. From the question, our **target** (Success given/under condition X) is 542. Thus, the conditional rate rate(Success | X) is calculated by:

$$\text{rate(Success} \mid \text{X)} = \frac{542}{700} = 0.774 = 77.4\%$$

Remark. The notation rate(target | condition), is used often for the **conditional rate**. It basically means the **rate** (*target* given/under the *condition*), which implicitly specifies the **population** to be the data that satisfies the condition.

Application of conditional rate

One real-world application of the conditional rate is that it can give us information about the **association** between two variables. We will see more about this later in this chapter.

But for now, with our example in table 2.1, since $\text{rate(Success} \mid \text{Y)} = \frac{289}{350} = 82.6\%$ is higher than $\text{rate(Success} \mid \text{X)} = \frac{542}{700} = 77.4\%$, we can intuitively find out that Treatment Y is more effective than Treatment X.

2.1.3 Joint Rate

Definition 2.1.7 ▸ Joint-Rate

Joint Rate means that our **target** is **not a marginal data** and our **population** is the **whole population** (the right-bottom corner of the table 2.1).

Remark. A joint rate problem usually has the keyword **"and"** appearing in the question.

Example 2.1.8 ▸ Joint Rate Example

For example, with Table 2.1, if we want to know "what proportion of patients received Y treatment and had a failed treatment outcome?".
Solution: This is a **joint rate** problem due to the appearance of the keyword "and". Here, our **target** is the data that is both Y and failure, so it will be 61. And based on the nature of the joint rate, our **population** is the whole population, which is 1050. So, the joint rate can be calculated as follows:

$$\text{rate(Y and Unsuccessful)} = \frac{61}{1050} = 0.0581 = 5.81\%$$

2.2 Association

Definition 2.2.1 ▸ Association

In statistics, **association** describes a *relationship* between two categorical variables where the presence or value of one variable provides information about the probability of the categories of the other variable. When two variables are *associated*, they are not independent; their results are linked in some way.

More specifically, we have two types of association, here let's briefly get the idea of what they are:

1. **Positive association**, which indicates that the two positive associated variables are likely to occur together.
2. **Negative association**, which is stated as a **comparison** against the *positive association*, it does not necessarily mean that the two variables are opposites but rather that their tendency to occur together is weaker than in a positive association.

> *Remark.* Association ≠ causation. While we observe a relationship, other factors (e.g. kidney stone size, patient health) could influence the results.

Before we give out a mathematical method to determine whether an association is positive or negative, let's consider two characteristics within a population, labeled as A and B. Here, A represents the presence of a particular trait (with NA indicating its absence), and B represents another trait (with NB indicating its absence). For example, let A denote the smoking habit—so that individuals with A are smokers and those with NA are non-smokers. Similarly, let B denote gender — so that individuals with B are male and those with NB are female.

Thus, the main focus is to compare the rate of A among those with B (rate(A | B)) to the rate of A among those with NB (rate(A | NB)). For instance, if the proportion of smokers among males **differs from** that among females, it suggests an **association** between smoking and gender. Otherwise, there is **no association** between smoking and gender.

2.2.1 Positive Association

Definition 2.2.2 ▸ Positive Association

We say A is **positively associated** with B if rate(A | B) > rate(A | NB). This means that the presence of A **when B is present** is stronger compared to **when B is absent**.

2.2.2 Negative Association

Definition 2.2.3 ▸ Negative Association

Similarly, we say A is **negatively associated** with B if rate(A | B) < rate(A | NB). This means that the presence of A **when B is present** is weaker compared to **when B is absent**.

Besides the two rules in the definition, we have summarized the following rules which can also be used to determine whether an association is positive or negative. These rules are mathematically proven by the symmetric rule which we will cover later in this chapter.

Establishing association	
Positive association between A and B: (any of the following)	**Negative association between A and B:** (any of the following)
$\text{rate}(A \mid B) > \text{rate}(A \mid NB)$	$\text{rate}(A \mid B) < \text{rate}(A \mid NB)$
$\text{rate}(B \mid A) > \text{rate}(B \mid NA)$	$\text{rate}(B \mid A) < \text{rate}(B \mid NA)$
$\text{rate}(NA \mid NB) > \text{rate}(NA \mid B)$	$\text{rate}(NA \mid NB) < \text{rate}(NA \mid B)$
$\text{rate}(NB \mid NA) > \text{rate}(NB \mid A)$	$\text{rate}(NB \mid NA) < \text{rate}(NB \mid A)$

Table 2.2: Positive/negative association

> *Remark.* In positive/negative association decision, for now, always remember what your A is and what your B is. Don't mix them with NA or NB. Otherwise, likely you will get the opposite result.

2.3 Two rules on rates

2.3.1 Symmetry Rule

Theorem 2.3.1 ▸ Symmetry Rule

The formal symmetry rule has three parts.
Symmetry Rule Part 1:

$$rate(A \mid B) > rate(A \mid NB) \Leftrightarrow rate(B \mid A) > rate(B \mid NA)$$

Symmetry Rule Part 2:

$$rate(A \mid B) < rate(A \mid NB) \Leftrightarrow rate(B \mid A) < rate(B \mid NA)$$

Symmetry Rule Part 3:

$$rate(A \mid B) = rate(A \mid NB) \Leftrightarrow rate(B \mid A) = rate(B \mid NA)$$

The three parts of the symmetry rules are similar. So, the proof for one part can be easily applied to the remaining. Here, let's prove the Part 1 logically.

Proof:

1. If rate(A | B) is more than rate(A | NB), then this means that there is a positive association between A and B
2. This means that we are more likely to see A when B is present, compared to when B is absent.
3. This is in turn means that we are more likely to see B when A is present, compared to when A is absent.
4. Hence rate(B | A) is more than rate (B | NA), proof!

Remark. The symmetry rule can also be proved mathematically. Here I will leave it for the readers.

2.3.2 Basic rule on rates

Theorem 2.3.2 ▶ Basic rule on rates

The overall rate(A) will always lies between rate(A | B) and rate(A | NB)

Consequence 1

$$\textit{The closer rate(B) is to 100\%, the closer rate(A) is to rate(A | B)}$$

Consequence 2

$$\textit{If rate(B)} = 50\%, \textit{then rate(A)} = \frac{1}{2}[\textit{rate(A | B)} + \textit{rate(A | NB)}]$$

Consequence 3

$$\textit{If rate(A | B)} = \textit{rate(A | NB)}, \textit{then rate(A)} = \textit{rate(A | B)} = \textit{rate(A | NB)}$$

Remark. For the basic rule on rates, note that rate(A | B) does not necessarily have to be the left bound. Instead, the smaller value should be the left bound whatever rate(A | B) or rate(A | NB).

To put it simply, the basic rule on rates states that the overall occurrence rate of event A in the whole **population** (**rate(A)**) is always a weighted average of **rate(A | B)** (A given B occurs) and **rate(A | NB)** (A given B doesn't occur). This means **rate(A)** cannot be higher or lower than both conditional rates — it lies between them.

The three consequences mainly states that

1. **Dominant Group Effect**: If almost everyone is in group **B** (e.g., 99%), **rate(A)** closely matches **rate(A | B)**. The rare "non-B" group has little influence.
2. **Balanced Groups**: If **B** and **NB** are equally common (50% each), **rate(A)** is the simple average: $\text{rate(A)} = \frac{\text{rate(A | B)} + \text{rate(A | NB)}}{2}$
3. **No B Influence**: If **rate(A | B) = rate(A | NB)**, then B doesn't affect A. The overall **rate(A)** equals both conditional rates, regardless of how common **B** is.

2.4 Simpson's Paradox

Definition 2.4.1 ▶ Simpson's Paradox

If we divide the whole population into several subgroups, we may find that a trend that appears in **more than half of the subgroups** of data may **disappears or reverses** when the subgroups are combined, this is called **Simpson's Paradox**.

Remark.
1. Here, "disappears" means the two variables in question (say A and B) are no longer associated, that is rate(A | B) = rate(A | NB).
2. The process of dividing the whole population into several groups is called *slicing*.

Example 2.4.2 ▶ Simpson's Paradox Example

Use the table 2.1 we have introduced at the beginning of the chapter. Let's divide the whole population into 2 subgroups using the third variable "stone size".

	Large stones			Small stones			Total (Large+Small)		
	Succ.	Total trt.	R(succ.) in %	Succ.	Total trt.	R(succ.) in %	Succ.	Total trt.	R(succ.) in %
X	381	**526**	**72.4%**	161	**174**	**92.5%**	542	700	77.4%
Y	55	80	68.8%	234	270	86.7%	289	350	**82.6%**

Figure 2.2: Slicing based on stone size

Here, we can observe that in both large size and small size group, the rate(Success | X) > rate(Success | Y). However, after we combine these two subgroups, rate(Success | X) < rate(Success | Y). Thus, a Simpson's Paradox has appeared! But why?

1. Let us look at the numbers highlighted in blue in the table. A crucial observation at this point is that treatment X seems to be used to treat mostly patients with large stones as compared to small stones. Thus, by the **basic rule on rates**, we know that the overall success rate of treatment X will be **closer** to the large stones success rate of 72.4% than the small stones success rate of 92.5%. Indeed, we have the overall treatment X success rate to be 77.4%.
2. Similarly, the overall success rate of treatment Y will be closer to the small stones success rate 86.7%. Indeed, the overall success rate of treatment Y is 82.6%.
3. Combining these two observations, it is no wonder that we have the overall success rate of X to be lower than the overall success rate of Y.

Conclusion: Slicing the data into the small and large stone subgroups reveals that

treatment X is indeed a better treatment!

2.4.1 A brief introduction to confounder

Let's take a brief look from the example above. In this example, we have introduced a third variable "large stone size", which is used to slice the whole original population. And based on some analysis, we will find out that the variable "large stone size" has association with **both** of the variables whose relationship we were initially investigating, thus affecting the conclusion of our initial study. Such a variable is called a *confounder*. We will discuss more about it in the later section.

Remark. When a Simpson's Paradox is observed, it implies that there is **definitely** a confounding variable present. However, the existence of a confounder does not necessarily lead to us observing Simpson's Paradox.

2.5 Confounders

Definition 2.5.1 ▶ Confounders

A **confounder** is a third variable that is **associated** with both the independent and dependent variables whose relationship we are investigating.

Remark. Note that we do not specify the direction (positive or negative) of association here. As long as the variable is associated in some way to the main variables, we will call it a *confounder*, or a *confounding variable*.

2.5.1 Application of confounders

The application of confounders can be summarized as follows:

1. **Importance of Data Collection**: Collecting data on **potential confounders** (e.g., stone size in a medical study) is critical to identify and adjust for hidden variables that distort associations between treatment and outcome. Without this data, confounding effects remain undetected, leading to biased conclusions.
2. **Practical Challenges**:
   - **Cost and Complexity**: Gathering additional variables is often resource-intensive.
   - **Analytical Overload**: Even with sufficient data, controlling for multiple confounders complicates analysis (e.g., "slicing" data across many variables).

3. **Limitations in Observational Studies**:
    - Observational studies (non-randomized) inherently face uncontrolled confounding because groups being compared (e.g., treated vs. untreated) differ in ways beyond the treatment itself.
    - Despite adjusting for known confounders, **residual confounding** (unmeasured or unknown variables) can persist.
    - Thus, observational studies can only establish association, not causation, as causality requires eliminating all confounder influence.

2.5.2 A solution to confounding

To solve the issue caused by the confounding variable, we can utilize the method of **randomization** (a.k.a randomized assignment). Why?

Recall that confounders arise when variables are **unequally distributed** between groups, leading to biased associations between the treatment and the outcome. For example, in the kidney stones study, stone size became a confounder because patients with larger stones were more likely to receive treatment X than treatment Y.

So, randomization is useful in this context because it ensures that both known and unknown confounders are, on average, equally distributed across all treatment groups. This balanced allocation eliminates the systematic relationship between any potential confounder and the treatment assignment, thereby isolating the treatment effect and minimizing bias in the results.

For example, if the allocation of large (and small) stone size cases to the two treatment types was **done randomly**, which tends to result in an equal proportion across the two groups, there would no longer be any association between stone size and treatment type. In this case, stone size would no longer be a confounder.

While randomization effectively balances confounders, it can raise ethical concerns. For instance, assigning treatments by a coin toss may compromise patient autonomy by removing their choice, which is critical in healthcare decisions. When ethical issues prevent true randomization, researchers must rely on alternative methods, such as adjusting for suspected confounders.

> *Remark.* Note that a confounding variable is associated to **both** the independent and dependent variables, so removing one of the associations is enough to remove the confounding variable.

3

Dealing with Numerical Data

3.1 Univariate EDA

Definition 3.1.1 ▶ Distribution

A **distribution** is an orientation of data points, broken down by their **observed number** or **frequency of occurrence**[a].

[a] That's why we can think distribution as a kind of frequency table

Before we go on, let's take a look at the dataset that we will be using in this chapter,

	A	B	C	D	E	F	G	H	I	J	K
1	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
2	1/1/2017	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44	Improved	1979	61 years 04 months	232000
3	1/1/2017	ANG MO KIO	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67	New Generation	1978	60 years 07 months	250000
4	1/1/2017	ANG MO KIO	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	67	New Generation	1980	62 years 05 months	262000
5	1/1/2017	ANG MO KIO	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1980	62 years 01 month	265000
6	1/1/2017	ANG MO KIO	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	67	New Generation	1980	62 years 05 months	265000
7	1/1/2017	ANG MO KIO	3 ROOM	150	ANG MO KIO AVE 5	01 TO 03	68	New Generation	1981	63 years	275000
8	1/1/2017	ANG MO KIO	3 ROOM	447	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1979	61 years 06 months	280000
9	1/1/2017	ANG MO KIO	3 ROOM	218	ANG MO KIO AVE 1	04 TO 06	67	New Generation	1976	58 years 04 months	285000
10	1/1/2017	ANG MO KIO	3 ROOM	447	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1979	61 years 06 months	285000
11	1/1/2017	ANG MO KIO	3 ROOM	571	ANG MO KIO AVE 3	01 TO 03	67	New Generation	1979	61 years 04 months	285000
12	1/1/2017	ANG MO KIO	3 ROOM	534	ANG MO KIO AVE 10	01 TO 03	68	New Generation	1980	62 years 01 month	288500
13	1/1/2017	ANG MO KIO	3 ROOM	233	ANG MO KIO AVE 3	10 TO 12	67	New Generation	1977	59 years 08 months	295000
14	1/1/2017	ANG MO KIO	3 ROOM	235	ANG MO KIO AVE 3	04 TO 06	67	New Generation	1977	59 years 08 months	295000
15	1/1/2017	ANG MO KIO	3 ROOM	219	ANG MO KIO AVE 1	07 TO 09	67	New Generation	1977	59 years 06 months	297000
16	1/1/2017	ANG MO KIO	3 ROOM	536	ANG MO KIO AVE 10	07 TO 09	68	New Generation	1980	62 years 01 month	298000
17	1/1/2017	ANG MO KIO	3 ROOM	230	ANG MO KIO AVE 3	04 TO 06	67	New Generation	1978	60 years	298000
18	1/1/2017	ANG MO KIO	3 ROOM	570	ANG MO KIO AVE 3	10 TO 12	67	New Generation	1979	61 years 04 months	3.00E+05
19	1/1/2017	ANG MO KIO	3 ROOM	624	ANG MO KIO AVE 4	04 TO 06	68	New Generation	1980	62 years 08 months	301000
20	1/1/2017	ANG MO KIO	3 ROOM	441	ANG MO KIO AVE 10	07 TO 09	67	New Generation	1979	61 years	306000

Figure 3.1: Singapore HDB resale flats within the period of Jan 2017 to June 2021

Remark. **distribution** is often used on **one variable**.

Example 3.1.2 ▶ Distribution Example

In this example, we want to investigate the distribution of the **Age** variable. So, we can get the following frequency table.

Age	Frequency
2	9
3	8
4	583
5	1105
6	884
7	295
8	255
⋮	⋮

Figure 3.2: Distribution of Age

However, looking only at the frequency table, it's hard to filter out the valuable information. So, here we introduce a more "graphical" way – histogram.

3.1.1 Histogram

Definition 3.1.3 ▸ Histogram

A **histogram** is a *graph* that groups data into **bins** (intervals) on the x-axis, with the y-axis showing the **frequency** or **count** of data points in each bin.

Example 3.1.4 ▸ Histogram Example

So, to put the frequency table we have in example 3.1.2 into a histogram, we can use either Excel or Radiant to generate the histogram as follows,

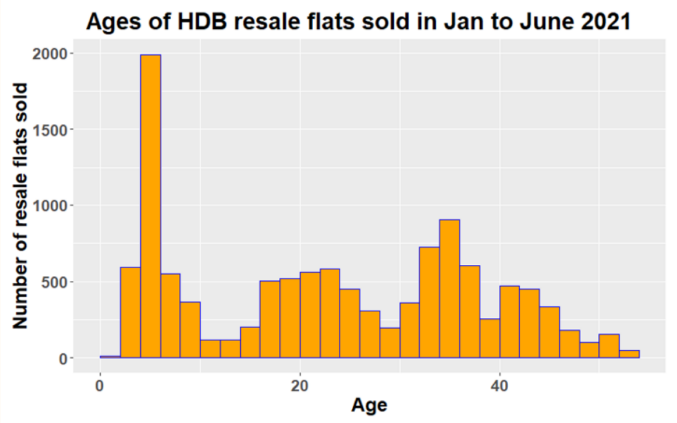

Figure 3.3: Histogram of Age

Once we have obtained a histogram, we are interested in its **overall pattern** and **deviation**

1. **Overall pattern**
   (a) Shape
   (b) Center

(c) Spread of the distribution

2. **Deviation**: identify the *outliers*

Shape

There are two important descriptors when we discuss the shape of a distribution, namely the *peaks* and the *skewness*.

1. **Peak**: Peak in a histogram highlight **the most common values** in specific intervals, which helps identify where the data is **most concentrated**. (We can think it as the local maximum in Math)
   - **Unimodal Distribution**: Has a *single*, distinct **peak**.
   - **Multimodal Distribution**: Contains **multiple** peaks.
   - **Bimodal Distribution**: A special case of multimodal distributions with exactly **two** peaks.
2. **Skewness**: For **unimodal distribution**, we can use another descriptor called *skewness* to describe the shape of the distribution. The following is three types of skewness.

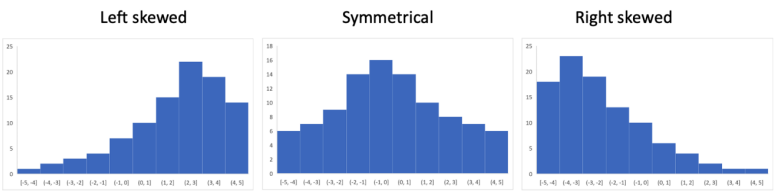

Figure 3.4: Skewness

   - **Normal Distribution**: The *normal distribution* is a famous **symmetrical** distribution, which is commonly known as the bell-curve.

Central

In this part, we are more interested in the **central tendency**, which refers to statistical measures like the mean, median, and mode that identify the central or most typical value in a data distribution. The following is the central tendency for the three skewness

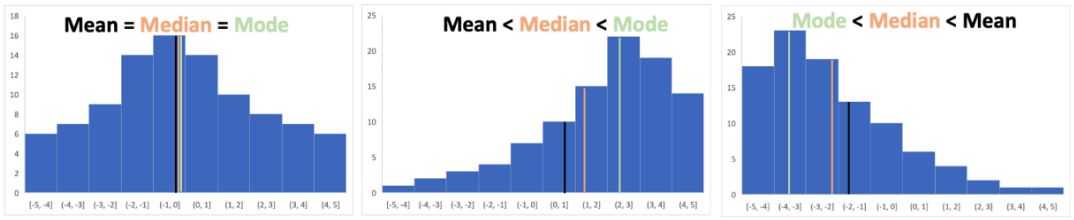

Figure 3.5: Central Tendency

Remark. These three rules usually hold, but **not always** hold.

Spread

This spread of a distribution refers to how the data **vary around the central tendency**. To study the spread, we will use **standard deviation** and **range**.

1. **Standard Deviation**: This has been introduced in the previous chapter.
2. **Range**: range = the largest data point − the smallest data point

Outlier

Definition 3.1.5 ▸ Outlier

An *outlier* is an observation that falls well above or below the overall **bulk** of the data.

Remark. Sometimes, we should not remove *outliers* because they do convey some useful information.

In general, we should keep the following tips when deciding the **bin width** of a histogram.

1. Avoid histograms with bin widths that are too large.
2. Avoid histograms with bin widths that are too small.
3. Our initial choice of bin width may not be the most appropriate.

Histogram vs. Bar Chart

Histogram:

- Depicts the distribution of a **numerical** variable along a continuous number line.
- Bars **touch** each other, reflecting adjacent numeric intervals.
- The order of bars **cannot** be changed because they represent a progression of numeric values.

Bar Chart:

- Displays data for **categorical** variables, where each bar corresponds to a distinct category.
- Bars typically have **gaps** between them.
- The order of bars can be **rearranged** freely without affecting the interpretation of the data.

3.1.2 Boxplot

Besides histogram, we can use **boxplot** to visualize the distribution of a numerical variable. To construct a boxplot, we need to following five numbers,

1. Minimum
2. Quartile 1(Q1)
3. Median (Q2)
4. Quartile 3(Q3)
5. Maximum

Furthermore, we also need to calculate the Interquartile range (IQR). While median can be viewed as the center of a data set, the IQR is a way to quantify the **spread** of a data set.

$$\text{IQR} = Q_3 - Q_1$$

After that, we can use a mathematical way to decide which points are *outliers*. This is done using the following theorem

Theorem 3.1.6 ▸ Find Outliers

A data point is considered an ***outlier*** *if it satisfies* ***one*** *of the following conditions:*

- *The value of the data point is* ***greater than*** $Q_3 + 1.5 \times IQR$
- *The value of the data point is* ***less than*** $Q_1 - 1.5 \times IQR$

To construct the boxplot, we can follow the 5 steps below

1. Draw a box from Q_1 to Q_3
2. Draw a vertical line in the box where the median (Q_2) is located
3. Identify all the outliers by using the theorem 3.1.6 above.

4. Extend a line from Q_1 to the smallest value that is **not** an outlier and another line from Q_3 to the largest value that is **not** an outlier. These lines are called *whiskers*
5. Mark each of the outliers with dots or asterisks.

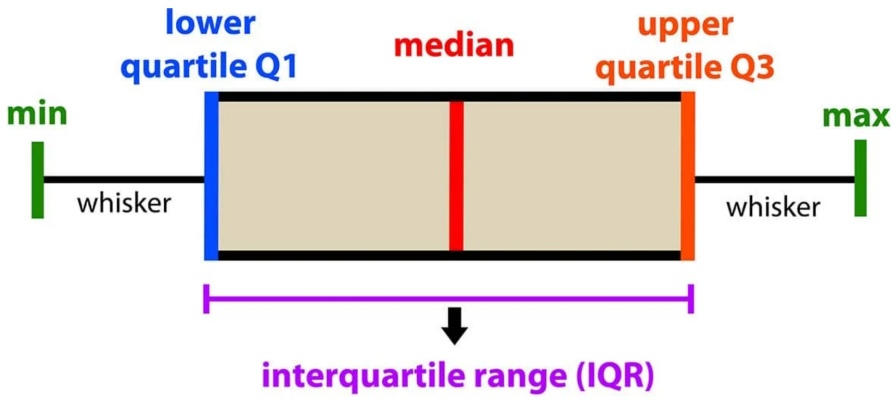

Figure 3.6: Boxplot Explanation

Boxplot vs. Histogram

1. A **histogram** typically gives a better sense of the *shape* of the **distribution** of a variable, compared to a **boxplot**.
2. If we wish to compare the **distributions** of different **data sets**, putting the different **boxplots** side by side is more illustrative than using **histograms**.
3. To identify and indicate **outliers**, **boxplots** do a better job than **histograms**.
4. The number of **data points** we have in a **data set** is better shown in a **histogram** than in a **boxplot**.

3.2 Bivariate EDA

In this section, we will be focusing on studying the **Bivariate**[1] **EDA**, especially the **relationship between two variables**. And based on the relationship, we can divide this section into two parts,

1. Deterministic relationship
2. Non-deterministic or statistical relationship

[1] Bivariate data is data involving two variables.

3.2.1 Deterministic

Definition 3.2.1 ▶ Deterministic

Deterministic means that the value of one variable can be **determined exactly** (this means that the determined value is unique) if we know the value of the other variable.

The most common type of deterministic relationship is the one that involves the conversion of units of measurement from one metric to another.

Example 3.2.2 ▶ Deterministic Relationship Example

For example, the relationship between Fahrenheit (F) and Degree Celsius (C) in the measurement of temperature is **deterministic**.
Solution: The relationship can be written as:

$$C = (F - 32) \times \frac{5}{9}$$

3.2.2 Non-deterministic

However, as we have said before, the main focus of this section is on a relationship between two variables that is not deterministic in nature. This is also more common in our daily life.

Definition 3.2.3 ▶ Non-deterministic

Non-deterministic[a] means that when given the value of one variable, we can only describe the **average value** of the other variable. And this kind of relationship is also called **association**.

[a] "Non-deterministic" is sometimes referred to as "statistical"

When doing the Bivariate EDA, we may find that unlike Univariate data, a frequency table, a boxplot or a hitogram is less useful for showing associations between two variables. So, in Bivariate EDA, we will mainly do,

1. **Scatter Plot**: For better visualization in Bivariate Data, a scatter plot is used to visualize the relationship between the two variables by displaying **data points**.
2. **Correlation Coefficient**: After examining the scatter plot, a correlation coefficient quantifies the **strength and direction** of any **linear association**.
3. **Regression Analysis**: Finally, regression analysis fits a line or curve through the data points, allowing for predictions of one variable based on the other.

Remark. Note the **regression** is used to predict, we **cannot** use the correlation coefficient to predict!

3.2.3 Scatter Plot

Definition 3.2.4 ▶ Scatter Plot

A **scatter plot** is a graphical tool that displays pairs of data values for two variables as individual points on a coordinate system. The overall pattern of these points helps reveal the type and strength of the **association** between the variables.

Remark. In Scatter plot, the x-axis is your **independent variable**, the y-axis is your **dependent variable**.

For example, we will be studying the **age** and **resale price** in Figure 3.1, and the following will be an example of what our scatter plot looks like,

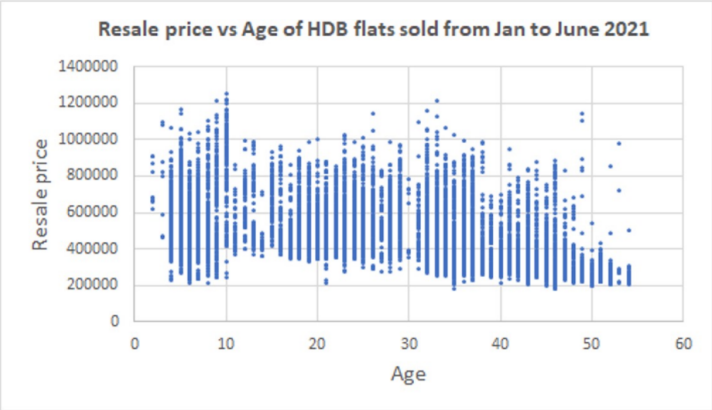

Figure 3.7: Scatter Plot Example

So, now how should we look at our scatter plot? Below is the table that summarizes the difference between what we should look from Univariate data and Bivariate data.

Univariate data		**Bivariate data**	
Overall pattern	**Deviation**	**Overall pattern**	**Deviation**
1) Shape 2) Center 3) Spread	Outliers	1) Direction 2) Form 3) Strength	Outliers

Table 3.1: Comparison of univariate and bivariate data characteristics

From the table above, we can clearly see that

1. **Univariate Data**: the key descriptors are **Shape** (symmetrical/skewed), **Center** (median, mean, mode), **Spread** (interquartile range, standard deviation, range).
2. **Bivariate Data**: the key descriptors are **Direction**, **Form**, **Strength** of the association.
3. **Common Aspect**: Data points that significantly deviate from the main pattern are **outliers**.

Direction

Definition 3.2.5 ▸ Direction

Direction describes the nature of the relationship between two variables: a **positive** relationship means that increases in one variable correspond to **increases** in the other; a **negative** relationship means that increases in one variable correspond to **decreases** in the other; if neither pattern is observed, the relationship has no specific direction.

For example, below is an example of the **direction** in the scatter plot.

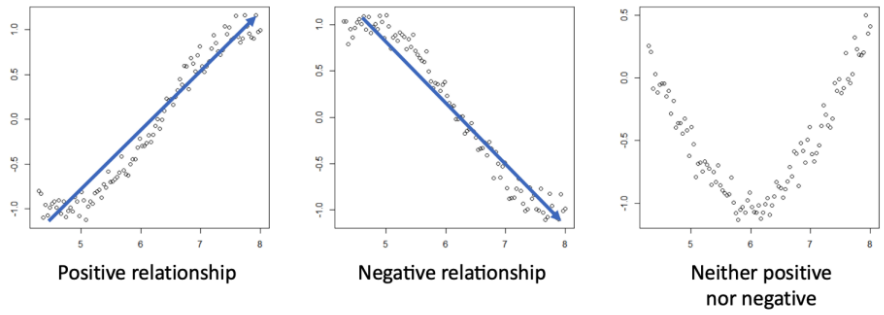

Figure 3.8: Scatter Plot Direction

Remark. Not all relationships can be classified as either positive or negative and there are those that do not behave in one way or the other.

Form

Definition 3.2.6 ▸ Form

The form of the relationship describes the overall shape of a scatter plot. It is classified as **linear** when the data points cluster around a straight line, and as **non-linear** when they form a smooth curve (e.g., quadratic or exponential patterns).

For example, below is an example of the **form** in the scatter plot.

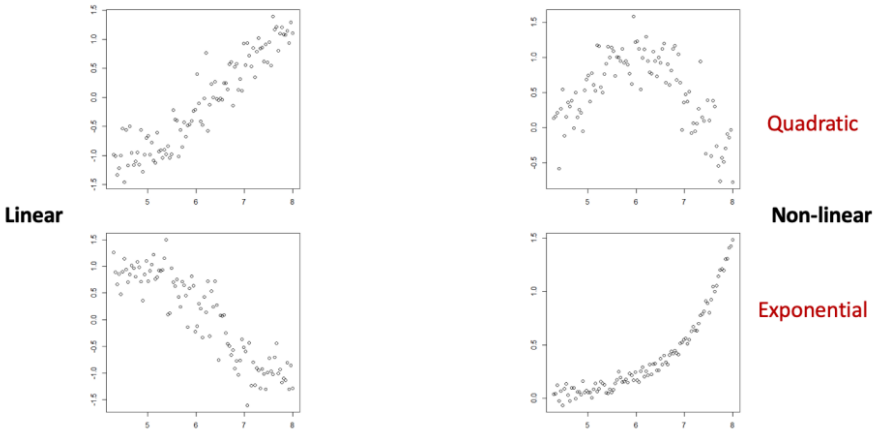

Figure 3.9: Scatter Plot Form

The two scatter plots on the left shows a linear form of the relationship between the two variables while the two scatter plots on the right shows non-linear forms

Strength

Definition 3.2.7 ▶ Strength

The **strength** of the relationship indicates **how closely** the data follow the **form** of the relationship.

For example, below is an example of the **strength** in the scatter plot.

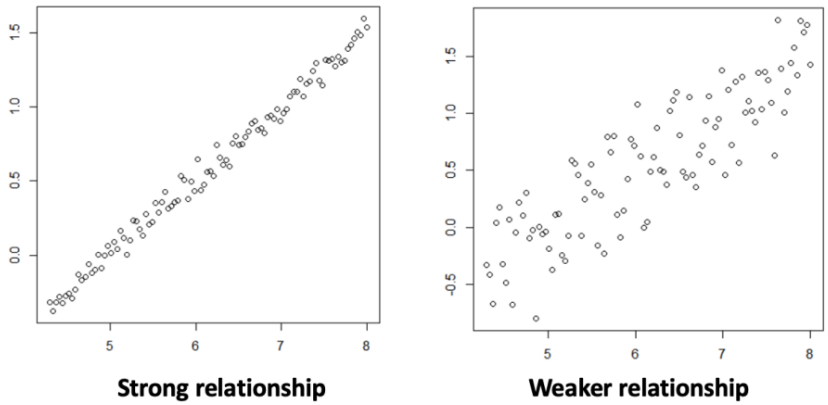

Figure 3.10: Scatter Plot Strength

3.2.4 Correlation Coefficient

Definition 3.2.8 ▶ Correlation

The **correlation coefficient** between two numerical variables is a measure of the **linear** association between them. The correlation coefficient, denoted by r, always ranges between -1 and 1.

We can use r to summarize the **direction and strength** of linear association between two variables.

1. If $r > 0$, the association is **positive**: an increase in one variable is associated with an increase in the other.
2. If $r < 0$, the association is **negative**: an increase in one variable is associated with a decrease in the other.
3. $r = 1$ indicates **perfect positive association**; $r = -1$ indicates **perfect negative association**.
4. $r = 0$ signifies **no linear association**.
5. The **sign** of r reveals the direction, while the **magnitude** (how close r is to 1 or -1) indicates the **strength** of the association.

Remark. **No linear association** between variables does not necessarily mean no association between variables. The relationship can be **non-linear** also, like the form of **quadratic** we have seen previously.

When describing the **strength** of a linear relationship, we usually follow the rule of thumb as given in the diagram below,

Figure 3.11: Correlation Coefficient Rule of Thumb

Compute the correlation coefficient r

The first method is to compute r **by hand** (**not recommended**)

To do so, let's use the table that shows a total of 10 data points of Bivariate data (x, y)

x	9	4	5	10	6	3	7	2	8	1
y	41	17	28	50	39	26	30	6	4	10

Table 3.2: Sample (x, y) data

1. **First compute the mean and standard deviation of x and y**: For this dataset, we find the mean and standard deviation of x to be 5.5 and 3.03 respectively while the mean and standard deviation of y are 25.1 and 15.65 respectively.

2. **Convert each value of x and y into *standard units***: To convert x (resp. y) into its standard unit, we compute
$$\frac{x-\bar{x}}{s_x} \text{ (resp.} \frac{y-\bar{y}}{s_y}) \tag{3.1}$$
where s_x and s_y are the standard deviations of x and y respectively. The table 3.3 below shows the values of x and y after they have been converted to standard units.

x	1.16	-0.50	-0.17	1.49	0.17	-0.83	0.50	-1.16	0.83	-1.49
y	1.02	-0.52	0.19	1.59	0.89	0.06	0.31	-1.22	-1.35	-0.96

Table 3.3: Standardized (x, y) values

3. **Compute the product xy in their standard units for each data point**: The table 3.4 below has an additional row for the value xy for each data point.

x	1.16	-0.50	-0.17	1.49	0.17	-0.83	0.50	-1.16	0.83	-1.49
y	1.02	-0.52	0.19	1.59	0.89	0.06	0.31	-1.22	-1.35	-0.96
xy	1.17	0.26	-0.03	2.36	0.15	-0.05	0.15	1.41	-1.11	1.43

Table 3.4: Values of x, y, and their product xy

4. **Sum the products xy obtained in the previous step over all the data points and then divide the sum by $n-1$, where n is the number of data points. The result is the correlation coefficient r**: For the data set above,
$$r = \frac{1}{9}(1.17 + 0.26 - 0.03 + 2.36 + 0.15 - 0.05 + 0.15 + 1.41 - 1.11 + 1.43) = 0.64$$

Remark. For the purpose of this course, you are not required to calculate the correlation coefficient by hand. However, by knowing this calculation method, you may develop some basic intuition on the properties of r

The second method is to use the calculator to get the correlation coefficient r.
(**Recommended**)

Properties of correlation coefficient r

Theorem 3.2.9 ▶ Properties of Correlation Coefficient r

1. *r is **not** affected by **interchanging** the x and y variables.*
2. *r is **not** affected by **adding** a number to **all** values of a variable.*
3. *r is **not** affected by **multiplying** a **positive** number to **all** values of a variable.*

Explanation:

1. **Interchanging x and y does not affect r**
   - The formula for r is:
   $$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x} \times \frac{y-\bar{y}}{s_y}\right)$$
   - Since multiplication is commutative, swapping x and y does not change the sum, meaning r remains the same.
2. **Adding a constant to all values of x or y does not affect r**
   - Adding a constant c shifts the mean by c, but does not change the standard deviation s_x or s_y, since it only affects location, not spread.
   - The standardized values $(x-\bar{x})/s_x$ remain unchanged, so r remains the same.
3. **Multiplying all values by a positive constant does not affect r**
   - Scaling x or y by a positive factor k scales both the numerator and denominator in standardization:
   $$\frac{kx-k\bar{x}}{ks_x} = \frac{x-\bar{x}}{s_x}$$
   - Since standard units remain unchanged, the correlation coefficient r remains unaffected.

Remark.

1. **Association is not causation**: With r being close to 1 or -1, we can only say there is a **strong association** between these two variables, and this is a *statistical relationship* between x and y instead of a *causal relationship*.
2. r **does not tell us anything about non-linear relationship**: r only measures the degree of **linear association** between two variables. For example,

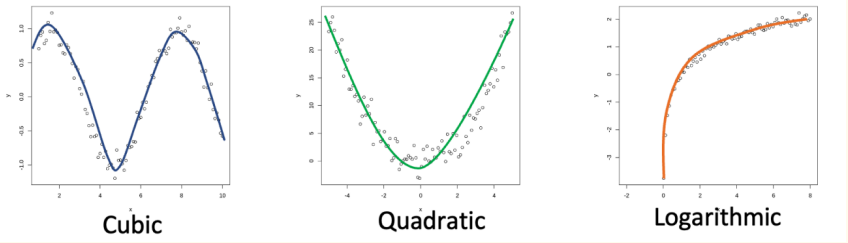

Figure 3.12: Non-linear association

The correlation coefficient r for these three scatter plots above are small but yet there is actually a **strong** relationship between variables. The value of r is small because the relationship between the variables is **not a linear one**.

> *Remark.* It is always a good practice to look at a scatter plot of the data set and not just deduce any relationship between the variables from the computed value of r.

3. **Outliers can affect the correlation coefficient significantly**

Ecological Correlation

Definition 3.2.10 ▸ Ecological Correlation

The **ecological correlation** represents relationships observed at the **aggregate level**, considering the characteristics of **groups** rather than **individuals**.

Think of it this way: when we compute correlation for **individuals**, each data point captures **personal variability**. But when we group individuals into **aggregates** (e.g., city averages instead of personal incomes), we **smooth out individual differences**. This reduction in variability makes patterns seem **stronger** than they actually are. The correlation appears inflated because it reflects group trends rather than true individual relationships.

Definition 3.2.11 ▸ Ecological Fallacy

The **ecological** fallacy occurs when incorrect inferences about individual-level behavior are made based on aggregate-level data.

For example, the following graph shows an **ecological fallacy**. In this figure, although the aggregate subgroup averages (denoted by four red points) show a **positive** correlation, the individual-level data within each subgroup actually reveals a weak **negative** correlation. Thus, assuming a positive individual correlation based on the aggregate data would be incorrect.

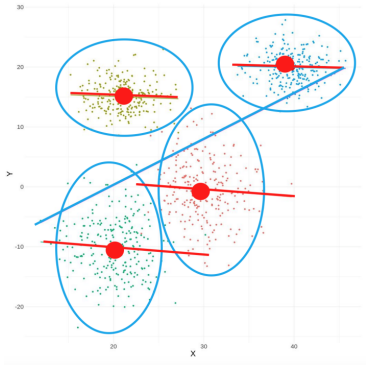

Figure 3.13: Ecological Fallacy

Definition 3.2.12 ▸ Atomistic Fallacy

The **atomistic fallacy** occurs when incorrect inferences about aggregate-level behavior are made based on individual-level data.

For example, the following graph shows an **atomistic fallacy**. From the individual-level data within each subgroup, we can see a clear positive relationship. But if we look at the aggregate subgroup averages (denoted by three red points), there is actually no relationship! Thus, an atomistic fallacy occurs!

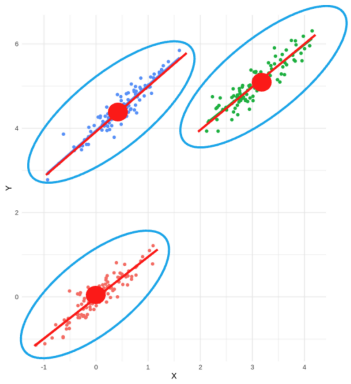

Figure 3.14: Atomistic Fallacy

To sum up, the following table summarizes the difference between ecological fallacy and atomistic fallacy.

Fallacy	Using	To Conclude
Ecological	Ecological correlation (aggregate level)	Individual level correlation
Atomistic	Individual level correlation	Ecological correlation (aggregate level)

Table 3.5: Ecological and Atomistic Fallacies

Remark. The aggregate level is the **subgroup average**, which is one point that represents its subgroup. The individual-level represents the **individual point within each subgroup**, **not** individual point within the **whole group**!

3.2.5 Linear Regression

Definition 3.2.13 ▸ Linear Regression

In this course, Linear regression is a statistical method used to model the relationship between the dependent variable and the independent variable by fitting a **linear equation** ($Y = mX + b$) to observed data.

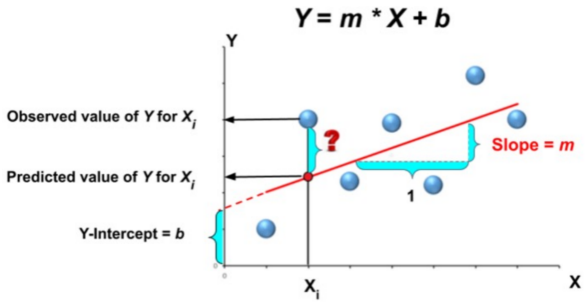

Figure 3.15: Linear Regression

Definition 3.2.14 ▸ Residual

From the figure 3.15, we define the **residual**[a] of the i-th observation as the observed value of Y for X_i (that is, Y_i) minus the predicted value of Y of X_i (predicted by the straight line).

[a] This residual, denoted by e_i, is sometimes called the **error** of the i-th observation as it measures how far the predicted value is from the observed value.

The Least Squares Method

The least squares method is a technique for finding the **best-fitting** line to a set of data points.

1. **Motivation:**
   - The goal is to minimize the discrepancy between the observed and predicted values.
   - We want to find the line that best approximates the data points by reducing the error.
2. **Formal Definition:**
   - Let the data points be $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
   - The model is $y = ax + b$, where a is the slope and b is the intercept.
   - The error for each data point is $e_i = y_i - (ax_i + b)$.
   - The least squares objective function is the sum of squared errors:

$$S(a, b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

3. **Calculation:**

- To find the optimal values of a and b, take the partial derivatives of $S(a, b)$ with respect to a and b.
- Set the partial derivatives equal to zero and solve the resulting system of equations.

> *Remark.* This step is usually down by the calculator or computer.

4. **Why Square the Errors?**
   - Squaring the errors ensures all errors contribute **positively** to the total sum.
   - Without squaring, large positive and negative errors could cancel each other out, making the total error misleading.
   - Squaring penalizes large deviations more heavily, resulting in a better and more robust fit.

Remark.

1. **Always pass through the average point**: The least squares regression line obtained from a set of observed data points (x_i, y_i) will **always** pass through that point of averages for that dataset, that is $(\bar{x}, \bar{y})$
2. **Not interchangeable**: We cannot change the dependent variable and independent variable.
3. **Difference between correlation coefficient r**:
   - The linear regression line is a **line** while the correlation coefficient r is just a **number**. (a.k.a to predict a value based on another value, we need **linear regression** instead of correlation coefficient.)
   - The gradient of the regression line m is **different from** the correlation coefficient r. But they share the **same sign** because of ths following relationship between them
     $$m = \frac{s_Y}{s_X} r$$
     where s_X (resp. s_Y) is the standard deviation of X (resp. Y).
4. **Make prediction only within the range of dependent variable**

Non-Linear Relationship

Some non-linear relationship can be converted to become **linear**. For example, an exponential relationship

$$y = cb^t$$

where c and b are some constants that we will determine. It can be converted to a linear relationship by taking the ln operation on both sides

$$\ln y = \ln(cb^t) \equiv \ln y = \ln c + t \ln b$$

Thus, if there is an exponential relationship between y and t, then we would expect to see a **linear relationship** between $\ln y$ and t. And the following are the steps to determine c and b

1. For each data point (t, y), compute $(t, \ln y)$
2. Find the linear regression line for $\ln y$ vs. t
3. Use the gradient and y-intercept to find the value of c and b
4. Rewrite the final exponential relationship

4

Statistical Inference

4.1 Statistical Inference

Definition 4.1.1 ▶ Statistical Inference

Statistical inference refers to the use of **samples** to draw inferences or conclusions about the **population** in question.

Our final goal in data analysis is to generate some information about the **population**. However, from the previous chapter, we only do EDA on the **sample**. So, we are more interested in **whether similar conclusions made at the sample level can be made at the population level also**. The following graph shows how statistical inference fits into exploratory data analysis (EDA) framework.

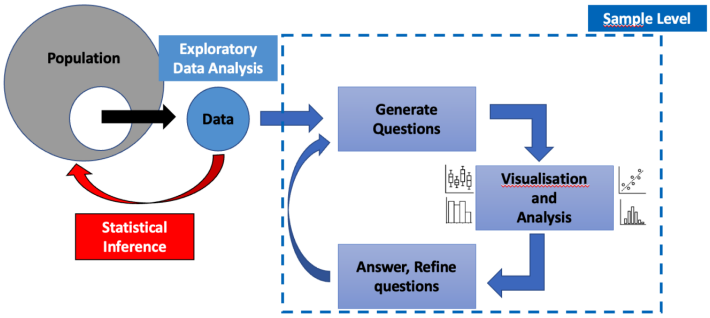

Figure 4.1: Statistical inference and EDA

Recall that the population parameter is a numerical fact about the **population**. When we take a sample from the population, the use of a sample statistic to estimate the population parameter is subjected to inaccuracies, which typically include *bias* and *random error*.

$$\text{Sample statistic} = \text{population parameter} + \text{bias} + \text{random error}$$

First thing first, in order to use our sample to make inference about the population, the **fundamental rule for using data for inference** should be met.

Theorem 4.1.2 ▶ Fundamental Rule for using Data for Inference

*Available data can be used to make inferences about a much larger group if the data can be considered to be **representative** with regards to the question of interest.*

By adopting good sampling method (e.g. using simple random sampling) and practices (e.g. having a good sampling frame), selection bias can be reduced. In addition, having a high response rate will minimise non-response bias. If bias can be reduced to an **insignificant level**, this would allow us to say

$$\text{Sample statistic} = \text{population parameter} + \text{random error}$$

The quantity of **random errors** refers to small differences that arise as a result of the sampling variability when using any probability-based sampling method.

> *Remark.* In the following content, we assume that **bias** can be reduced, so we only care about the random errors.

In this course, the statistical inference includes two types

1. confidence intervals
2. hypothesis testing

4.1.1 Confidence interval

Definition 4.1.3 ▶ Confidence Interval

A **confidence interval** is a **range of values** that is likely to contain a population parameter based on a certain degree of confidence. This degree of confidence is called the **confidence level** and is usually expressed as a percentage(%)

For this course, we will only introduce the construction of confidence intervals for population **proportion** and **mean**.

Confidence Interval for population proportion

The formula we use is

$$p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}} \tag{4.1}$$

where

$$p^* = \text{sample proportion}$$
$$z^* = \text{“z-value” from standard normal distribution}$$
$$n = \text{sample size}$$

Remark. The z^* is usually chosen by the software. If we want to calculate manually, then we have

$$\text{For a 90\% confidence interval, } z^* = 1.645$$
$$\text{For a 95\% confidence interval, } z^* = 1.96$$

To interpret a confidence interval, we mainly have to look at two parts (suppose that our 95% confidence interval is 0.254 ± 0.0191)

1. The **confidence level** (for example, 95% in our example); and
2. The interval (0.254 ± 0.0191 in our example)

The value 0.0191 is known as the *margin of error* which directly impacts the width (how wide/narrow) of the confidence level.

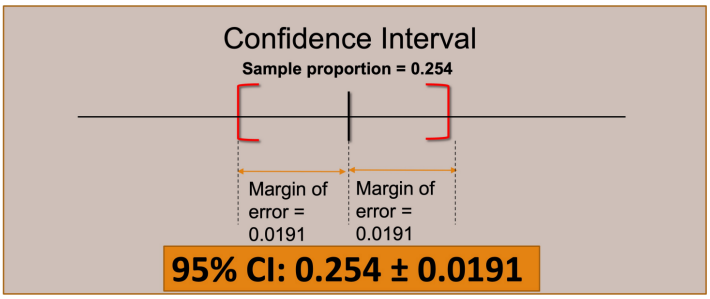

Figure 4.2: Confidence Interval on Population Proportion Example

So, the interpretation we have may be as follows:

We are 95% confident that the population proportion (the parameter in this case) lies within the confidence level.

Using the idea of repeated sampling [1], the interpretation of “95% confident” is that if **many simple random samples of the same size are taken, and a confidence level is con-**

[1] It means that many simple random samples of the same size are taken and with the different sample statistics obtained from the different samples, different confidence intervals are constructed using the same method as above.

structed for each of them, then **about 95% of the confidence intervals constructed would contain the population parameter**.

This means if we collected 100 simple random samples and their 95% confidence intervals were computed in the same manner, then about 95 out of 100 confidence intervals will contain the population parameter.

> *Remark.* It is wrong to say that there is a 95% chance that the population parameter is **in the interval** (and 5% chance that it is not)! The element of chance (or probability) comes from the **uncertainty of sampling** rather than the **uncertainty in the value of the population parameter**, whereas the latter is always fixed!

Properties of Confidence intervals

1. The larger the sample size n, the smaller the random error (a.k.a margin error).
2. The higher the confidence level at which the confidence interval is constructed (a.k.a the larger the z^*), the wider the confidence interval.

Confidence Interval for population mean

The formula we use is

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}} \tag{4.2}$$

where

$$\begin{aligned} \bar{x} &= \text{sample mean} \\ t^* &= \text{“t-value” from t-distribution} \\ s &= \text{sample standard deviation} \\ n &= \text{sample size} \end{aligned}$$

> *Remark.* The exact value of t^* depends on the sample size n and the confidence level of the confidence interval we are constructing. Again, this is chosen by the software.

For example, let’s suppose the confidence interval for the population mean we found to be 448727 ± 6706.01, we can similarly get an illustrative graph as follows,

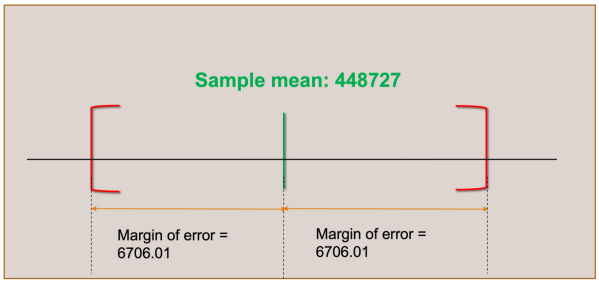

Figure 4.3: Confidence Interval on Population Mean Example

4.1.2 Hypothesis Testing

In statistics, when we want to test an idea or make a decision about something (like whether a new drug works or if a coin is fair), we use a method called **hypothesis testing**. This involves setting up two competing statements: the **null hypothesis** and the **alternative hypothesis**.

Definition 4.1.4 ▶ Hypothesis Test

A **hypothesis test** is a statistical inference method used to decide if the data from a random sample is **sufficient** to support a particular hypothesis about a population.

In this course, we will focus on two types of hypothesis about the population, in particular, whether,

1. a population parameter is x;
2. in the population, 2 categorical variables A and B are associated with each other.

Definition 4.1.5 ▶ Null Hypothesis

Null hypothesis H_0 is the default assumption.

1. In case 1[a], it says there's **no effect, no difference**, or that things are equal to a specific value.[b] We assume it's true unless we find strong evidence against it.
2. In case 2[c], it means there is **no association**[d] between the two categorical variables.

[a] population parameter is x
[b] Think of it as the "nothing special is happening" statement.
[c] association between two categorical variables
[d] Independence

Definition 4.1.6 ▶ Alternative Hypothesis

Alternative Hypothesis H_1 is the opposite — it's what we're trying to find evidence for.

1. In case 1, it says there is **an effect, a difference**, or that something doesn't equal a specific value.[a]
2. In case 2, it means there is **an association**[b] between the two categorical variables.

[a] It's the "something interesting is happening" statement.
[b] Dependence

Remark. The **null** and **alternative** hypotheses should be mutually exclusive, meaning that they **cannot be true simultaneously**.

Imagine you're testing a new teaching method to see if it improves students' test scores compared to the old method.

- **Null Hypothesis** (H_0): The new teaching method doesn't make a difference—the average test scores are the same as with the old method.
- **Alternative Hypothesis** (H_1): The new teaching method does make a difference—the average test scores are higher (or different) with the new method.

Five steps for hypothesis testing

1. Identify **the question and state the *null hypothesis* and *alternative hypothesis***.[2]
2. Set the ***significance level*** of our test.[3]
3. Using our sample, we find the relevant sample statistic.[4]
4. With the sample statistic and the hypothesis, we can calculate the p-value.[5]
5. Make a conclusion of the hypothesis test.[6]
   (a) If the p-value is **smaller than** the significance level, you **reject the null hypothesis** and say there is evidence for the alternative hypothesis.
   (b) Otherwise, you **fail to reject the null hypothesis**, meaning you don't have enough evidence to support the alternative.

[2] How these hypotheses are stated depends on the context of the question and our aim.
[3] For this course, you just need to know it's about choosing a number.
[4] This means calculating the population parameter we want but using the sample data. In **association tests**, this means we assume the null hypothesis is true and recalculate the corresponding data.
[5] This is usually done by the software.
[6] What the conclusion turns out to be depends on the p-value calculated and the significance level set for the test.

Note: You don't "accept" the null, you just don't have proof against it.

Definition 4.1.7 ▶ *p*-value

The *p*-value is the probability of obtaining a result **as extreme** or **more extreme** than our observation in the direction of the alternative hypothesis, **assuming the null hypothesis is true**.

Remark. This *p*-value is usually calculated by the software.